Heterogeneous Effects in Education: The Promise and Challenge of Incorporating Intersectionality into Quantitative Methodological Approaches*

Lauren Schudde

The University of Texas at Austin

In preparation for Review of Research in Education

January 7, 2018

Heterogeneous Effects in Education: The Promise and Challenge of Incorporating Intersectionality into Quantitative Methodological Approaches

Various dimensions of individuals' identities—including race, class, gender, and sexuality—overlap and intersect. Intersectionality posits that rather than operating as "mutually exclusive entities," the dimensions combine to produce an individual's experience and shape broader social inequalities (Collins, 2015, p. 2). Most extant literature on intersectionality relies on a limited range of methodological approaches, partially due to the complexity of the concept (McCall, 2005). To date, the theory of intersectionality has largely guided qualitative efforts in social science and education research. Limiting inquiry on intersectionality to a narrow set of methodological tools restricts the scope of knowledge on the subject (McCall, 2005, p. 1722). Translating the construct to new methodological approaches is inherently complex and challenging, but offers the possibility of breaking down silos that keep education researchers with similar interests—but different methodological approaches—from sharing knowledge.

Quantitative approaches that emphasize the varied impacts of individual identities on educational outcomes move beyond singular dimensions capturing individual characteristics, drawing a parallel to intersectionality. Scholars interested in *heterogeneous effects* (HE) (also referred to as differential, conditional, marginal, and heterogeneous treatment effects) recognize the shortcomings of focusing on the effect of a single social identity. HE allows scholars to examine how various identities, skills, and positions affect the experience and outcomes of individuals (Elwert & Winship, 2010; Museus & Griffin, 2011; Rhodes, 2010; Turney, 2015). This integrative review explores techniques used in quantitative research to examine differential effects across individual background, drawing on methodological literature from the social

sciences and education. I examine the goals and challenges of approaches to capture HE and explore how they relate to intersectionality by leveraging McCall's (2005) examination of intersectionality. I conclude by discussing what education researchers can learn from other applied fields that are working to develop a crosswalk across the two disparate, but interconnected, literatures.

Bringing HE and Intersectionality into Conversation

Individuals' experiences and responses to those experiences vary across individuals based on who they are (Berger & Luckmann, 1966; Elwert & Winship, 2010). HE, like intersectionality, anticipates that background characteristics, skills, and group memberships influence individuals' experiences and the effects of those experiences. Despite commonalities in the assumptions and goals of intersectional and HE research, this alignment has received little attention. At best, mixed methods researchers acknowledge that quantitative methods can benefit from intersectional theory, but emphasize the role interview data could play in illuminating patterns of individual and group experiences that are difficult to capture through large-scale survey data alone (Griffin, Bennett, & Harris, 2011). Nevertheless, many quantitative researchers in the field of education rely on secondary survey data, which leaves them with little to no influence on sampling strategy and survey items. To understand the impact of student identity, including membership in multiple groups, on outcomes of interest, researchers must have a way to account for intersectional identities without collecting additional data.

The literature on intersectionality in education and the social sciences does not fully explore the "wide range of methodological approaches" that can be used to study "multiple, intersecting, and complex social relations" (McCall, 2005, 1773). McCall (2005) describes common approaches to capturing complex social relationships, including the approach most

widely used in quantitative scholarship. Leveraging existing analytical categories to examine relationships and shifting dynamics of inequality, quantitative approaches capture intersectional identities and related "intercategorical complexity" by using pre-defined social groups (McCall, 2005, p. 1773). Although less holistic than qualitative approaches that deconstruct analytical categories and less critical than approaches focused on identities that "[cross] the boundaries of traditionally constructed groups," the intercategorical approach allows researchers to draw upon large data sets to examine the interplay of pre-defined identity characteristics (e.g., Black, female, homosexual) in producing unequal outcomes (McCall, 2005, pp. 1773-1774). The intercategorical approach also has the capacity to use large-scale data and evaluate the generalizability of results, which can counterbalance its limitations.

While much of the quantitative literature fails to recognize variation in individual responses as intersectionality, many applications exploring HE are squarely focused on interpreting the way that individuals react differently to a similar environment or experience. The work examining effect heterogeneity[1] acknowledges the way that various identities, programs, and environments interact to "shape complex social inequalities" (Collins, 2015, p. 2). In other words, extant quantitative research does intersectional work (though it is limited in the specific intersections observed through existing categories), but the language differs.

Applied researchers who wish to examine effect heterogeneity should do so with great intentionality, developing theory to inform their hypotheses about varied returns to specific experiences (e.g., Bedeian & Mossholder, 1994; Elwert & Winship, 2010). Incorporating concepts from intersectionality can inform the quantitative literature. Leveraging theory from intersectionality could also improve researchers' ability to interpret HE results, often presented

[1] Throughout the piece, I alternate between referring to the exploration of variation in effects as effect heterogeneity and heterogeneous effects, or HE.

in the form of statistically significant interactions between two or more variables. However, I found several logistical barriers to integrating an intersectional framework into quantitative approaches, including limitations of existing data that lead to difficult tradeoffs in research design and challenges in interpreting complex results.

Drawing on McCall's (2005) work as a theoretical framework for examining the quantitative approaches to capture effect heterogeneity, I examined the methodological literature with an eye toward how extant quantitative approaches can be used to capture intersectionality, how intersectionality can inform future quantitative work, and how extant data (including nationally representative survey data and administrative data) align with the goals of intersectionality. This paper proceeds as follows: First, I describe the methods I employed for the literature review, including inclusion criteria for studies selected for the review. Next, I synthesize the methodological literature on approaches to capture HE, challenges for implementation, and considerations to overcome those challenges. I conclude by discussing the segmentation of the education literature by methodological approach and the role that intersectionality can play in bridging the divide between quantitative and qualitative scholarship and in advocating for more theoretically grounded statistical models.

Strategy for Literature Review

To find relevant literature on heterogeneous effects and understand the current state of methodological practice and challenges, my primary goal was to identify methodological literature on HE from social science and education. The methodological literature on HE offers a means to examine the underlying assumptions and goals of the approaches as they relate to intersectionality. I used EBSCOhost to search for *heterogeneous effects*, as well as alternative forms of the term (*heterogeneous treatment effects*, *differential effects*, *conditional effects*,

marginal effects), and methodological terms associated with the approach: *modified regression* and *interaction term*. I narrowed the search to work published from 1989 (the year of Crenshaw's first publication on intersectionality) to 2017 and allowed for journals, working papers, and reports. The initial EBSCOhost search offered 2,952 search results from EconLit, 2,227 from ERIC, and 3,588 from SocINDEX.

I worked to narrow the results based on the following inclusion criteria. Given my focus on methods to examine heterogeneous effects, the first inclusion criterion required that research focus on methodological approach—e.g., how to obtain heterogeneous effects, difficulties in interpretation and challenges in application, and how to improve the statistical approach—rather than application alone. The second inclusion criterion required that papers describe HE in the context of using observational data, as the intercategorical approach to intersectionality largely relies on pre-determined categories available in existing data (McCall, 2005, p. 1787).

The majority of the search results were applications of regression using interaction terms, rather than discussions of the approach, its challenges, and interpretations. It was not feasible to identify (and eliminate) applications through search terms. I was able to winnow the results to further align with my focus on observational data by adding a "NOT" Boolean operator for the terms "randomized controlled trial" and "meta-analysis." The new search offered 2,948 from Econlit, 2,164 from ERIC, and 3,538 from SocINDEX on which I performed a title review. To sort through the search results, I first examined titles, keeping only papers that did not explicitly note a particular population/dataset/application. When I reached 500 consecutive entries that did not meet the inclusion criteria, I stopped the search. The search based on title resulted in 51 unique papers after sorting through the first 2,100 results (I found 55 papers total, but four were working paper drafts of published versions already included).

For the remaining 51 papers, I examined the abstract and manuscript to determine if they met the inclusion criteria. Upon closer review, several did not discuss estimating heterogeneous effects, but how to produce estimates when individuals received varied treatments (i.e., where individuals are exposed to different levels of an intervention)—thus they were not focused on effect moderation, but, rather, different treatments (n=11). Several other papers were too narrowly focused on the application to fulfill the first criterion (n=7). After the full review, I identified 32 for inclusion in the synthesis.

I gathered additional literature on methodological approaches through ancestor searches, adding papers based on the reference lists of those in my initial search (n=14). Finally, I incorporated papers based on my background knowledge, primarily to describe additional implications of the statistical approaches or support the discussion of intersectionality (n=24). To offer additional insights on the methodological approaches, I incorporated examples from NCES data documentation, literature on sampling weights, and information on power analysis. Sources that were not obtained through the literature search or ancestor search are denoted with an asterisk in the reference list.

I organized the synthesis of the literature into three themes: common statistical approaches, major challenges in application, and overcoming those challenges when using observational data to examine HE. Within the results, the most common approach for estimating heterogeneous effects uses interaction terms in regression analysis. A second approach extends the first by using propensity score strata, rather than individual covariates, to assess variation in effects. In leveraging these techniques, several challenges emerged in discussions regarding HE, including concerns over adequately supporting statistical models with theory, difficulty in interpreting the results, and the potential pitfalls of insufficient sample size to examine effects

among subgroups. I follow up with additional considerations for using existing largescale data and the role of administrative data as a potential means to overcome the challenges of small sample size.

Approaches to Capture Effect Heterogeneity

Intersectionality, starting with Crenshaw's (1991) work on Black women, focuses on the multidimensionality of an individual's experience and stands in contrast to a "single-axis analysis" that would otherwise distort those experiences (i.e., examining the experience of being Black or female disparately) (p. 139). Most social scientists, particularly those interested in inequality, agree that responses to experiences vary across individuals and between groups (Elwert & Winship, 2010; Kam & Trussler, 2007; Manski, 2007; Wodtke & Almirall, 2015; Xie, 2011; 2013). For education researchers, we might anticipate variation in students' responses to a given educational experience based on their background (i.e., certain background characteristics modify the impact of the experience on the measured outcome). Modified regression and heterogeneous treatment effects across propensity scores are two statistical approaches that allow researchers to explore variation in individual responses. I describe the two approaches below.

Modified Regression

Regression analysis estimates the relationship between covariates (also referred to as independent variables or predictors) and the outcome (the dependent variable). A hypothesis of differential effects anticipates that a *moderator*, or more than one moderator, influences the strength of the relationship between two other variables. When the effect of a given variable depends, in some way, on the value of another variable, there is an *interaction* between the two variables (VanderWeele & Knol, 2014b). In a regression, the role of moderator—a covariate that may dampen or amplify the effect of another variable—is typically captured through a

multiplicative interaction term. The magnitude of the relationship between the independent variable of interest and the outcome varies as a function of another predictor (Flanders, DerSimonian, Freedman, 1992; Preacher, Curran, & Bauer, 2006; Wodtke & Almirall, 2015). In an educational setting the independent variable of interest, for example, might include exposure to an educational program, but it also could be a particular background characteristic or group membership. The presence of the interaction effect is typically evaluated by the statistical significance of the interaction term in the regression results. The interpretation of the interaction is critical and requires further investigation, which researchers often conduct by plotting and evaluating the slopes of different values of the modifying variable (for illustration, see Preacher et al. (2006)). The inclusion of interaction terms also renders the coefficient for the dependent variable more difficult to interpret (Flanders, DerSimonian, & Freedman, 1992).

Although interaction terms are simple to include in regression models, main-effects-only regression models are still the norm throughout social science and education research (Choo & Marx Ferree, 2010; Elwert & Winship, 2010; Rhodes, 2010; Turney, 2015). Most published research using regression accounts for individual background measures in predicting the outcome but fails to account for interactions between those measures. Yet understanding complex social processes requires "seeing and seeking complexity" when building statistical models, rather than starting with the simplest model (Choo & Marx Ferree, 2010, p. 146)
Why do researchers who theoretically believe in effect heterogeneity rely on main-effects only regression models? Elwert and Winship (2010) proposed that scholars assume that main-effects coefficients represent a "straightforward average" of heterogeneous individual level effects (p. 327). Researchers also rely on average effects because sample sizes may be too small to include interaction terms between the independent variable of interest and more than a few common

modifiers (gender, race, income, etc.) and the variables necessary to explicitly model heterogeneity remain unmeasured and/or unknown (p. 328). To illustrate that ignoring effect heterogeneity, as in most main-effects-only regressions, is prone to failure, Elwert and Winship (2010) used simulation, comparing results from models with unmodeled effect heterogeneity and results that capture effect heterogeneity. They found that unmodeled effect heterogeneity led to biased estimates (Elwert & Winship, 2010). Thus, capturing effect heterogeneity through interaction terms is important, but developing statistical models that "seek complexity" requires theoretical grounding, as I describe in the section on challenges in application.

Effect Heterogeneity Across Propensity Score Strata

A more recent approach for capturing variation in effects follows much of the same motivation as modified regression. Rather than including interactions between two or three measures to test for variation in the outcome, it leverages a composite of background characteristics and examines variation across the resulting score, referred to as a *propensity score*. A *propensity score model* estimates the predicted probability of participation using observed characteristics, summarizing that probability into one number (Hu & Mustillo, 2016; Morgan & Winship, 2007). While standard propensity score methods focus on average treatment effects, much like regressions without interaction terms, recent research explores HE by leveraging a "stratification multilevel model" (Hu & Mustillo, 2016, p. 71). To test for variation in results across the probability of experiencing a given treatment, scholars use propensity scores from the initial model to disaggregate effects of the treatment (Xie, Brand, & Jahn, 2012).

The approach outlined by Xie et al. (2012), which they call "Heterogeneous Treatment Effects" (HTE), examines effects across intervals of propensity scores. In addition to describing the technique, they offer a program, HTE, to execute the approach in Stata. HTE compares effect

sizes for those with the lowest probability of receiving treatment with those with increasingly higher probabilities of selection by dividing the propensity score distribution into strata (i.e., treatment*propensity to participate in treatment, where the propensity scores are divided into intervals). Analyzing the pattern of treatment effects as a function of the propensity score (i.e., do students with a higher propensity for selection benefit more than those with a lower propensity?) has to potential to uncover the "implications of the distribution of social resources, policy interventions, and events across the population" (p. 320).

Rather than examining how one background factor or identity moderates the outcome, researchers can use this method to understand how individuals' backgrounds—including the composite intersecting identities—influence selection into treatment and variation in effects. For example, Brand and Xie (2010) used data from the National Longitudinal Survey of Youth 1979 (NLSY79) and the Wisconsin Longitudinal Study to model college students' propensity to earn a bachelor's degree. The propensity score model included several background measures such as race, parents' income and educational attainment, gender, high school class rank, and cognitive ability. The scholars then examined how economic returns to a bachelor's degree varied across students' propensity to earn a degree. By relying on a summary measure of pretreatment characteristics, the HTE approach avoids exhausting precious degrees of freedom compared with testing an array of interaction effects across individual covariates (as in modified regression) (Turney, 2015). Using a hierarchical linear model with students nested in propensity score strata, Brand and Xie (2010) examined the pattern of effects on earnings. They found a statistically significant negative pattern of effects across propensity score (students with the lowest probability of completing college demonstrated the biggest returns for earning a degree), which

Brand and Xie interpreted as evidence that those who are least likely to earn a degree benefit the most from doing so.

Despite the novelty and advantages of HTE, scholars using the approach face some challenges in interpreting results. Individuals within each propensity score stratum do not share the same exact intersecting identities. Descriptive statistics of each propensity score stratum can demonstrate the most common identities in the group. In Brand and Xie's (2010) study, male college students in the stratum with the lowest propensity to complete college were disproportionately non-White and grew up in households with relatively low parental income and educational attainment compared to the sample average (p. 286, Table 3). However, because the stratum includes various intersections of identities, it is difficult to summarily conclude which students (based on particular identifies) are most likely to benefit from a degree.

For the purpose of modeling intersectionality, the propensity score method is more flexible than the modified regression approach because it captures various intersecting identities in one interaction term. For the same reason, however, it is less intuitive for interpreting the implications of the results for specific subgroups. While the approach has parallels with the goals of intersectionality, Brand and Xie (2012) did not invoke intersectionality or explicitly consider it when discussing the implications of their results. Doing so may have offered them additional language with which to describe the composition of students in specific strata and the complex pattern of effects.

Further Approaches to Examine Effect Heterogeneity

Because effect heterogeneity is "endemic to nearly all social contexts," capturing that variation can offer valuable insights for social theory and inform program and policy implementation (Wodtke & Almirall, 2015, p. 3; Xie et al. 2012). As such, there are additional

statistical approaches to assess HE, depending on the research question and data available. Through path analyses, researchers can examine moderation by incorporating interaction terms, in addition to exploring mediation (Fairchild & MacKinnon, 2009; Henseler & Chin, 2010). Other approaches, like an instrumental variable approach, align with causal inference, but allow interaction terms to test for effect heterogeneity (Heckman, Urzua, & Vytlacil, 2006; Moffitt, 2008). Generally, these approaches grapple with many of the same challenges as those in modified regression when incorporating interactions into the models. For additional information, the citations above offer some insights on challenges specific to each approach.

Challenges in Application

From a methodological standpoint, it would not be terribly difficult for more scholars to include interaction terms in their regression models (Elwert & Winship, 2010; Franzese & Kam, 2009; Rhodes, 2010). Yet, there are several notable challenges to doing so. In this section, I highlight four main challenges: supporting models with theory, examining tradeoffs in research goals when determining whether to examine HE, complex interpretation, and identifying data with adequate sample size to explore hypothesized interactions.

Supporting Statistical Models with Theory

Theory is a vital component of the process of building statistical models with interaction terms, whether scholars leverage modified regression or HTE. Rather than encouraging the inclusion of interaction terms in search of significance (sometimes referred to as "data snooping"), the methodological literature encourages strong theoretical justification for the statistical models (Aiken & West, 1991; Bedeian & Mossholder, 1994; Bobko & Russell, 1994, p. 194; Elwert & Winship, 2010). Using theory to inform model building is necessary to understand the need for and the interpretation of interactions. Work from qualitative research and

extant intersectional theory can guide quantitative researchers as they build models and accumulate evidence regarding the role intersecting identities play in how individuals' respond to experiences (Green, Evans, & Subramanian, 2017; Ragin & Fiss, 2016; Turney, 2015). This approach is especially valuable in applied fields like education, where the impact of programs and environments for different types of students has implications for practice and policy.

To align with intersectional theory, quantitative researchers using modified regression may wish to include multiple interactions terms and leverage three-way interactions when supported by theory (e.g., race*gender*class rather than just race*class). With more interactions (and, thereby, intersections), the model is better able to account for variation across identities. To date, it appears that most papers with interaction terms in education still focus primarily on two-way interactions, which offer insight into HE but often across singular dimensions of student background. The difficulty in incorporating three-way interactions may partially be driven by the difficulty justifying the inclusion of three-way interactions with minimal prior research to cite as an example. McCall (2005) argued that the evaluation of multiple interaction effects using intersectional theory may be discouraged by academic journals because reviewers often stress the need to cite already developed bodies of research and because editors pressure authors to cover more material in less space, which makes it difficult to theoretically justify a large number of interaction terms (McCall, 2005, p. 1787). Qualitative scholarship may be able to provide theoretical justification for exploring various interactions, as qualitative work often offers detailed information on participants' background. Even if faced with conflicting findings across extant qualitative research, quantitative researchers could leverage that as evidence that additional assessment is necessary (Turney, 2015).

Like modified regression, the HTE approach is best suited to research questions that anticipate variation in outcomes based on individual background, where participants in a given program/experience may receive different benefits based on how likely they were to participate. Efforts to examine HTE have sparked debate among researchers interested in causal inference and effect heterogeneity. Breen, Choi, and Holm (2015) argued that evidence of heterogeneous effects may actually be attributed to selection bias, cautioning researchers using the method on observational data collected in social settings. Leveraging the same nationally representative data as Brand and Xie (2010), they demonstrated how, in the presence of additional selection bias (e.g., a variable left out of the model) or a competing differential effect (where some students in the higher strata benefit more from college than their peers with lower propensities), it is possible to artificially identify a differential effect. Their critique illustrates the need for a rich set of covariates, detailed consideration of potential confounders, and careful exploration of alternative explanations for differential effects. To effectively leverage HTE, education researchers should closely consider why variation across propensity to participate might be present in response to the independent variable of interest and test alternative explanations using the data. By leveraging additional theory and examining competing hypotheses, the HTE approach offers insights that may help researchers understand whether and how groups of students respond differently to a given educational treatment.

Tradeoffs: Competing Goals in Applied Research

In applied, policy-relevant research there is an inherent interest in variation in effects across different groups of individuals. But applied work often navigates a tension between the need to examine differential effects and the desire to offer simple population-level statistics (see Morabia, 2014; VanderWeele & Knol, 2014a, 2014b). Education scholars may learn lessons

from a recent debate in epidemiology. Epidemiologist are increasingly interested in using interaction terms to discern whether some individuals stand to benefit more from an intervention than others—a pressing need in the face of limited resources (a challenge similarly faced in education) (VanderWeele and Knol, 2014b). These questions are important for programmatic decisions and implementation, but scholars face a parallel incentive to simplify results, presenting broad patterns of population-level trends rather than complex narratives (Morabia, 2014). Focusing on average effects moves scholars away from a complex vision of individuals and masks the way in which background characteristics and prior experiences predispose them to do better or worse than peers.

Yet even when epidemiologists include interaction terms in regressions, they often fail to include more than one modifier variable, despite theory that would support additional model complexity (Morabia, 2014). Minimizing complex statistical models save researchers from a "plethora of interactions" that could "render population thinking and group comparisons essentially useless" (VanderWeele & Knol, 2014b, p. 79). To consider the tradeoffs between exploring HE and relying on average effects, VanderWeele and Knol recommended that scholars, from the outset of their research, evaluate their goals and purpose in building a model that includes interactions. Do they seek to understand variation in effects more broadly? Do they seek to target certain subpopulations to determine how to maximize the effectiveness of an intervention or how to uncover mechanisms for improving its effectiveness? Whereas the first goal would leverage interaction terms for descriptive purposes, the second fuels evaluation of which subgroup to treat or how to be most effective in the face of limited resources. Researchers must consider the theory and goals driving their research as they build their statistical models

and identify the best approach. The inclusion of interaction terms must align with the overarching goals of the project.

Interpreting Complex Interaction Effects

The inclusion of interaction terms inherently makes models more difficult to interpret. While the incorporation of moderators in a statistical model should be supported by theory, theory does not necessarily make the output—which now includes a series of main-effects and interactions—easier to interpret. If researchers rely solely on regression coefficients, they may find it difficult to produce concrete and straightforward interpretations of the results. Results can be made more concrete—and easier to interpret—by computing predicted values for specific subgroups of individuals (Long & Freese, 2006).

Recent updates in statistical software have improved the tools available to help scholars interpret complex interaction effects (Jann, 2013; Williams, 2012). In 2011, Stata incorporated a new set of commands to help researchers produce predicted probabilities of specific subgroups that can be applied to interpret interactions. The margins and contrast commands increase the ease with which users can compute the predicted probability for a given hypothetical individual. Williams (2012) produced illustrative examples of the interaction of female*age to calculate the predicted probability that men and women will end up with diabetes (p. 318). Marginsplot, which helps build visuals from the interactions, increases the ease with which researchers can illustrate interactions (Williams, 2012). Jann (2013) leverages margins and marginsplot to show how users can illustrate varied patterns of effects across subgroups based on the interactions included in their models. The Stata command produced by Xie et al. (2012) relies on similar calculations across propensity score strata as the margins command and produces parallel graphics to marginsplot, facilitating the interpretation of HE across propensity score stratum.

While recent advances in statistical software increase the ease with which researchers can test and display specific interactions, the advice to rely on theory to guide statistical models still holds. Though HTE and marginsplot produce figures to illustrate variation in effects, the interpretation of those results relies on the researchers' knowledge of the literature. Extant research on intersectionality may be useful to inform the interpretation of complex interactions across multiple categories.

Sample Size and Statistical Power

When incorporating interaction terms into a model, researchers may find that very few individuals fall into certain categories (e.g., very few Black students are present in a given school). Small subgroup sample sizes can make it difficult to run the analysis or detect an effect. This challenge may contribute to the lack of three-way interactions in the literature. Since a three-way interaction (e.g., race*parent education*gender) requires even smaller subgroups of individuals in overlapping categories, identifying the impact, even if the intersection is potent for the outcome, may be more difficult. Failing to find a statistically significant impact that is otherwise present is referred to as a *Type II error*. While the hypothetical model may better account for variation across identity (if supported by theory, compared with a model that incorporates a two-way interaction), the number of students in each "combined" group of identities may be small, putting strain on the model, and resulting in the omission of some combined groups in the results.

Similar to standard modified regression, the HTE approach also requires attention to subgroup sample size—in this case, the propensity score strata. This problem is slightly less concerning in HTE than in modified regression because the approach relies on summary scores, minimizing the reliance on subgroups based on one specific covariate (Turney, 2015). However,

researchers may still find small numbers of individuals with estimated propensity scores in some strata. Hu and Mustillo (2016) provide a review of recent developments in propensity score methods and provide practical tips to evaluate sensitivity to sample size within strata and methods to adjust the number of strata (p. 74).

Given the emphasis on statistical significance in publishing, small cell size has important implications. For researchers interested in publication, the risk of failure to detect an effect is a powerful disincentive to pursue research questions that would rely on small subgroups (which have a greater risk for Type II error). There is a tradeoff between the potential contribution of including interaction terms and the limitations of secondary data to capture moderating relationships, largely due to sample size. Bobko and Russell (1994) noted the importance of considering statistical power for examining group-level difference early in research development. Ideally, this notion would arise in study design (i.e., data collection), but researchers using secondary data may also want to maintain this consideration during early analytic planning. If the power is insufficient to study an important subgroup or phenomenon across a given set of identities, researchers might revise their plans (Bobko & Russell, 1994). In some cases, researchers can raise their threshold for considering an effect statistically signficant, above the typical p-value of .05 (Marshall, 2005). Publically available computer programs[2] can help scholars estimate the power of their planned modified regression to detect a hypothesized effect. Such programs use empirically based algorithms to allow researchers to estimate statistical power by providing values for factors known to affect power, such as anticipated magnitude of the moderating effect and sample size of moderator-based groups (Aguinis, Beaty, Boik, & Pierce, 2005). These resources are quite valuable for researchers interested in examining

[2] Programs to detect statistical power for various forms of modified regression are available at: http://www.hermanaguinis.com/mmr/index.html. Instructions for use are available in Aguinis (2004).

intersectionality using subgroups in observational data, but otherwise uncertain "how small is too small?" in regard to sample size.

Of course, assessing the appropriate size of a sample to merit exploring interactions is more complex than simply adhering to power analysis results. Even with the risk of nonsignificant results, exploring HE can still be valuable (Bobko & Russell, 1994; Vandenbrouke, 2013; VanderWeele & Knol, 2014b). Effect heterogeneity often provides new insights compared to the alternative option of assuming average effects. The information gleaned can offer insights into the responses of narrow subpopulations—we can learn a lot from the pattern of results, even if it is unlikely that we will find statistically significant results. Performing a power analysis is one way for researchers to be informed about the sample size that would be required to identify a given effect. What is most important is that scholars feel confident that they have enough students in the subsample of interest to believe that the patterns reflect general trends, rather than idiosyncrasies in a tiny subgroup of sampled students.

This set of considerations aligns well with the themes in the intersectionality literature. There is value in examining theory-driven variation in student responses to programs and experiences, even if doing so does not return statistically significant results. Testing for interactions is about scientific reasoning and theory. There is no ideal outcome of the analysis, at least in terms of intellectual curiosity (though publishing bias leans toward statistical associations)—we should ask the question if answering it could bolster or refute theory (Morabia, 2014; Vandenbrouke, 2013).

Overcoming Challenges: Is Large-Scale Data Ready for Intersectional Analysis?

Related to the challenges noted above, I describe the current rationale and approach for drawing samples in National Center of Education Statistics (NCES) data sets. I consider whether

the sampling strategy aligns with the examination of group-level differences and interactions, which is necessary to leverage intersectionality. I also describe sampling weights and their limitations in overcoming the problem of small sample size for subgroups and examine the possibility of "big data," including state administrative data, in providing adequate sample sizes to conduct intersectional analyses.

NCES Data: Implications of Sampling Design for Intersectional Inquiry

Collecting large-scale survey data with the goal of achieving a nationally representative sample of students is challenging. In education, there is no comprehensive list from which to draw a random sample from the target population, whether it be kindergartners, high school sophomores, or first-time college students (Thomas & Heck, 2001). Even with a hypothetical list in hand, a random sample could not ensure that students with certain characteristics would be adequately represented, yet this representation is particularly important for researchers interested in specific subgroups of students and intersections with other identities (Thomas & Heck, 2001, p. 519). NCES addresses these issues with a multistage cluster sampling strategy, which involves oversampling students based on characteristics pre-determined to need additional representation in the sample, such as racial minorities (i.e., some individuals have a higher probability of selection) (Tourangeau et al., 2009).

Each NCES study has its own design in drawing a complex multistage sample. For instance, the Early Childhood Longitudinal Study (kindergarten cohort) (ECLS-K) follows a sample of kindergarteners. To select a nationally representative sample of kindergarteners in 1998-99, NCES started with a list of counties or groups of counties (Tourangeau et al., 2009). After selecting geographic areas, NCES selected schools within the region, and then children from within the schools. Not all children had an equal probability of being selected. To obtain

ensure precise estimates, NCES oversampled Asian and Pacific Islanders (Tourangeau et al., 2009, p. 4-2). As additional waves of data were added to the ECLS-K 1998-99, new priorities for subsample preservation arose and NCES adjusted the probability of selection (Tourangeau et al., 2009, p. 4-1). NCES indicates that oversampling was crucial to achieve adequate numbers of underrepresented subgroups of students—in this case, Asian and Pacific Islanders in wave 1 and transfer students and language minorities in wave 3 (Tourangeau et al., 2009).

Overall, it is important to note that NCES data, like most nationally representative data, are not designed specifically for intersectional data analysis. Restricting the sample to focus on certain students, for instance, in investigating the Black–White test score gap using the ECLS-K, has its limitations. Breaking down those subgroups even further may result in challenges such as low cell size. Although NCES oversamples racial minority students to improve researchers' ability to study certain subgroups of students, NCES would likely need to oversample additional underrepresented groups (e.g., stratifying the sample based on sexuality, disability status, or other background characteristics) to ensure adequate representation for intersectional analyses.

NCES has not published any reports of power analyses conducted to evaluate adequate sample size for student subgroups prior to data collection, though the language in their ECLS-K technical report suggests that some sort of analysis was conducted to determine the appropriate sample size for Asian and Pacific Islanders (Tourangeau et al., 2009, p. 4-4). Ideally, as research demands change, the sampling design and data structure will shift to allow for adequate representation of various other groups of students. As of right now, researchers interested in leveraging existing data must contend with the limitations of the data. In the next section, I examine the role of sampling weights and whether they alleviate concerns about small sample size when using nationally representative data.

Accounting for sampling design. Multi-stage sampling strategies yield samples that include disproportionate numbers of some individuals and do not align with subgroup representation in the population. Results that fail to adjust for sampling strategy are biased, where the extent of bias varies based on how the researchers restricted their analytic sample and selected variables (Thomas & Heck, 2001, p. 520). Researchers must address issues related to oversampling, where students have unequal probabilities of selection, and clustering, where students within some groups are more similar than those across groups. Thomas and Heck (2001) recommended that researchers using complex sample data incorporate either design-based strategies (e.g., sampling weights to account for selection probabilities) or model-based strategies (e.g., models, such as multilevel models, that account for clustering) into their research design.

Leveraging sampling weights. Sampling weights are used to align the sample's distribution for a set of variables with the population from which the sample was drawn (Winship & Radbill, 1994, p. 240). Oversampling based on racial identification—as in the ECLS-K—may result in a sample with a higher percentage of some students than is proportionally present in the population. In this case, sampling weights can be constructed to adjust the distribution toward what it would have been, had Asian students not been oversampled. The trouble arises when research moves beyond descriptive statistics—which many social scientists and education researchers aim to do—because sampling weights bias the estimation of standard errors (Solon, Haider, & Wooldridge, 2015).

Winship and Radbill (1994) and, more recently, Solon et al. (2015) acknowledged the bias produced by using sampling weights in multivariate statistical procedures such as regression. Many pressing and important inquiries in applied research focus on estimating statistical associations and cause-and-effect, rather than population descriptive statistics. Both

sets of approaches described to estimate HE aim to allow researchers to understand the impact of independent variables on an outcome and examine modifiers of that effect. When the goal is to understand the impact of one variable (or many variables) on another, then the use of sampling weights may not be appropriate.

While many researchers rely on weighted regressions to draw population-level inferences (indeed, some of the literature supports this method, e.g.: Aiken & West, 1991; Overton, 2001), weighted regression estimates are often less precise—they have larger standard errors—than unadjusted regressions (Dickens, 1990; Solon et al., 2015; Winship & Radbill, 1994). The problem appears to be due to the assumption that individuals' error terms are independent of one another, when they likely have group-level factors in common that are not accounted for by the weights (Dickens, 1990; Solon et al., 2015).[3] Decisions are also made in the research process that render the use of sampling weights less applicable. For instance, the researcher may narrow the sample in a way that makes it difficult to know how representative it is of the population.

Overall, researchers must use caution when relying on sampling weights to overcome sampling design decisions. The purpose of sampling weights is not to address small sample sizes, but to adjust the descriptive statistics of the sample to resemble the population. As such, using sampling weights does not resolve the problem of low statistical power. For the purposes explored in this paper, scholars interested in understanding educational impacts among certain subgroups or in examining group-level differences may consider using unweighted analyses (if they restrict their sample in a way that makes population inferences unnecessary) or using a strategy such as multilevel modeling to control for clustering related to the sampling design. I

[3] For more in-depth information on the origin of bias due to sampling weights, see Dickens (199), Solon et al. (2015), and Winship and Radbill (1994).

elaborate more on this issue next. If nothing else, researchers using the national data should consider the implications of sampling design when formulating their analytic plan.

Accounting for clustering. As noted in the previous section, sampling weights can account for oversampling. However, they do not account for clustering where perhaps some students in the sample share similar characteristics because they are from the same geographic region or attend the same school. Further, sampling weights must be used at a one level of analysis; the researcher must focus on either students or schools, rather than studying both levels of analysis simultaneously (Thomas & Heck, 2001; Winship & Radbill, 1994). Multilevel approaches take clustering into account by decomposing estimates for each variable into the part contributed from within a group/cluster—the individual student—and the part due to variation between clusters—often the schools from which the students were sampled.[4] Another strategy includes using robust standard errors adjusted to account for clustering. Adjustments can be performed easily in most statistical software. For instance, in Stata, researchers can use the vce(cluster) option while performing regressions (Stata, 2017).[5] This approach can be combined with sampling weights, if the researcher has not narrowed the sample in a way that makes sampling weights inappropriate.

"Big Data": State Administrative Data and Other Large Data Sources

Large-scale survey data may be underpowered for some intersectional analyses, making it difficult to identify HE across subgroups with small sample sizes. Is the solution to find "bigger" data? While increasing the sample size for NCES data is unlikely, due to resource constraints, more researchers are turning to administrative data to capture entire populations of individuals (Card et al., 2010). Relying on administrative data has its pros and cons. While the data are more

[4] For more information on multilevel modeling, see Muthen and Satorra (1995); Raudenbush and Bryk (2002).
[5] Cameron and Miller (2015) offer a useful overview on approaches to deal with clustering.

likely to be sufficiently powered to test interactions across multiple covariates, potentially detecting HE for various subgroups, the available measures tend to be limited (Card, Chetty, Feldstein, & Saez, 2010; Scott-Clayton & Wen, 2017).

Large-scale nationally representative surveys collect detailed self-reported data, which is sometimes combined with administrative data (e.g., transcripts), producing a rich set of covariates on which researchers can draw. In administrative data, the set of information for any individual in the population is finite. This is a problem for researchers interested in testing for variation in effects across various identities. Furthermore, statistical models missing covariates that influence on the outcomes may be biased (Cunha & Miller, 2014; Scott-Clayton & Wen, 2017).

Yet researchers have increasingly turned to administrative data to answer pressing policy problems, including in the field of education. While much of this research relies on state administrative data, recent research also leverages even larger data sets, including tax records (e.g., Chetty et al., 2014). Large sample sizes make the data sufficiently powered to test a variety of interactions, despite limitations in terms of the depth and breadth of available identity measures. This tradeoff means that some intersections can be explored, depending on the data source, but that researchers interested in examining intersectionality must consider which pre-determined categories are of interest and to pursue the data only if those categories are available. Access to administrative data sources can be more difficult to navigate and require a larger investment of money or time to obtain a data license than NCES studies (Card et al., 2010; Cunha & Miller, 2016).

Effect Heterogeneity and Intersectionality: A Path Forward

Intersectional theorists aim to overcome the tendency to "conflate or ignore intra-group differences" and variation in individual experiences across multiple identities (Crenshaw, 1991, p. 1241). Although intercategorical approaches for capturing intersectionality are limited in their ability to capture the complexity of individual experiences, examining HE offers the potential for researchers to illuminate intersectional effects across pre-defined groups (McCall, 2005). Increasing the role of intersectional analyses in quantitative research offers new means by which to examine variation in responses to lived experiences. Encouraging greater dialogue between the (mostly qualitative) scholarship on intersectional research and quantitative education research has the potential to improve theory formation for hypothesized interactions and to offer generalizable results using large-scale data.

Scholars increasingly acknowledge the need for overlap between quantitative approaches and intersectional theory. A recent issue of *Race, Ethnicity, and Education* explored whether quantitative methods can "support a critical race agenda in educational research" (Garcia, López, & Vélez, 2017, p. 2). In the issue, Gilborn, Warmington, & Demack (2017) take a skeptical stance, arguing that quantitative methods "cannot match qualitative approaches in terms of their suitability for understanding the numerous social processes that shape and legitimate …inequity" (p. 3). The authors acknowledge the role that quantitative methods play in highlighting structural barriers and inequalities faced by different groups of individuals, but warn that statistics often disguise inequities and protect the status quo. This skepticism highlights why researchers should be mindful of the purpose, design, and limitations of the data sets they use. They should also use caution in interpreting the lack of HE across subgroups as evidence of equal returns to the same experience or intervention; they must be mindful of the potential for type II errors.

In other applied fields, a growing subset of quantitative researchers explore how intersectionality and quantitative approaches can inform one another. Population health researchers (e.g., Bauer, 2014; Green et al., 2017) and poverty scholars (e.g., Ragin & Fiss, 2016) acknowledge the need to incorporate intersectionality into theory and interpretation of quantitative analyses. Green et al. (2017) argue that by combining intersectionality with other social theories related to the production of inequality in health outcomes, researchers are better poised to interpret interactions among measures of social identity as part of "interlocking systems of oppression" (pp. 215-216). Ragin and Fiss (2016) emphasize that only through considering a "combination of characteristics"—as opposed to the independent contribution of various independent variables—can researchers understand poverty and inform the complex policy changes to overcome it (p. 13). Population health and poverty research have several similarities to education, given both fields' interests in the effects of programs and policies on individual outcomes and the driving concern of how to improve outcomes among subgroups of individuals. Thus, these resources may be useful to education researchers to inform the field's conversation about intersectionality, pushing past traditional methodological divisions.

Conclusion: Are Intersectional and Quantitative Approaches Compatible?

Although large-scale data are not intentionally designed for conducting intersectional analyses, researchers use quantitative data to explore how individuals' characteristics, skills, and group memberships moderate responses to a variable of interest. In this paper, I synthesized the literature from quantitative methods to consider the merits and challenges of approaches available to explore heterogeneous effects consider the extent to which they align with intersectionality. I also explored the potential of new developments, whether it be statistical software or administrative data, to improve researchers' ability to incorporate intersectionality

into quantitative approaches.

There is tension between qualitative and quantitative research that becomes even more apparent in this line of inquiry. Quantitative work is criticized for taking on a deficit-based approach (Reid, Epstein, Pastor, & Ryser, 2000; Schreiner & Anderson, 2005). Meanwhile, qualitative scholars often must respond to critiques about sample size, generalizability, and the need for rigor. The disconnect between studies using quantitative techniques to examine effect heterogeneity and the primarily qualitative and theoretical literature on intersectionality comes as no surprise.

However, each side stands to be enriched by the other. Incorporating intersectionality would strengthen the toolkit available to researchers as they examine heterogeneous effects, supporting theory for statistical models and offering concrete examples from which to interpret results. Likewise, qualitative researchers could benefit from the capability of large-scale data to test the generalizability of their findings. Exploring patterns illustrated in the extant intersectional literature through quantitative data may bolster support for findings and pinpoint areas for inquiry in new contexts.

Innovations in statistical software and the availability of large-scale data make examining effect heterogeneity feasible for a broader array of researchers. The literature on methodological approaches emphasizes the need to leverage theory to support models that test for HE and to interpret the results. This paper represents an attempt to illustrate the overlap in interests among scholars studying effect heterogeneity and intersectionality. In the field of education, both lenses stand to provide "new angles of vision" to understand how practices, policies, and structures influence social inequality (Collins, 2015).

References[6]

Aiken, L.S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

Aguinis, H., Beaty, J. C., Boik, R. J., & Pierce, C. A. (2005). Effect size and power in assessing moderating effects of categorical variables using multiple regression: a 30-year review. *Journal of Applied Psychology, 90*(1), 94-107. doi:10.1037/0021-9010.90.1.94

Aguinis, H. (2004). *Regression analysis for categorical moderators*. New York, NY: Guilford Press.

Bauer, G. R. (2014). Incorporating intersectionality theory into population health research methodology: challenges and the potential to advance health equity. *Social Science & Medicine, 110*, 10-17. doi:10.1016/j.socscimed.2014.03.022

Bedeian, A. G., & Mossholder, K. W. (1994). Simple question, not so simple answer: Interpreting interaction terms in moderated multiple regression. *Journal of Management, 20*(1), 159-165. doi:10.1177/014920639402000108

*Berger, P. L., & Luckmann, T. (1966). *The social construction of reality: A treatise in the sociology of knowledge.* New York, NY: Doubleday.

Bobko, P., & Russell, C. J. (1994). On theory, statistics, and the search for interactions in the organizational sciences. *Journal of Management, 20*(1), 193-200. doi:10.1177/014920639402000111

Brand, J. E., & Xie, Y. (2010). Who benefits most from college? Evidence for negative selection in heterogeneous economic returns to higher education. *American Sociological Review, 75*(2), 273-302. doi:10.1177/0003122410363567

[6] I use an asterisk to denote references collected based on background knowledge, rather than the literature search strategy and ancestor search.

Breen, R., Choi, S., & Holm, A. (2015). Heterogeneous causal effects and sample selection bias. *Sociological Science, 2*, 351-369. doi:10.15195/v2.a17

*Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. Journal of Human Resources, 50(2), 317-372. doi:10.3368/jhr.50.2.317

*Card, D., Chetty, R., Feldstein, M., & Saez, E. (2010). *Expanding access to administrative data for research in the United States*. NSF SBE 2020 White Paper (September 2010). Washington, DC: National Science Foundation.

*Chetty, R., Hendren, N., Kline, P., & Saez, E. (2014). Where is the land of opportunity? The geography of intergenerational mobility in the United States. *The Quarterly Journal of Economics, 129*(4), 1553-1623. doi:10.1093/qje/qju022

Choo, H. Y., & Marx Ferree, M. (2010). Practicing intersectionality in sociological research: A critical analysis of inclusions, interactions, and institutions in the study of inequalities. *Sociological theory, 28*(2), 129-149. doi:10.1111/j.1467-9558.2010.01370.x

*Collins, P. H. (2015). Intersectionality's definitional dilemmas. *Annual Review of Sociology*, *41*, 1-20. doi:10.1146/annurev-soc-073014-112142

*Crenshaw, K. (1991). Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review, 43*(6), 1241-1299. doi:10.2307/1229039

*Cunha, J. M., & Miller, T. (2014). Measuring value-added in higher education: Possibilities and limitations in the use of administrative data. *Economics of Education Review, 42*, 64-77. doi:10.1016/j.econedurev.2014.06.001

Dickens, W. T. (1990). Error components in grouped data: Is it ever worth weighting? *Review of Economics and Statistics,* 328-333. doi:10.2307/2109723

Elwert, F., & Winship, C. (2010). Effect heterogeneity and bias in main-effects-only regression models. *Heuristics, probability and causality: A tribute to Judea Pearl,* 327-336. London, UK: College Publications.

Fairchild, A. J., & MacKinnon, D. P. (2009). A general model for testing mediation and moderation effects. *Prevention Science, 10*(2), 87-99.

Flanders, W. D., DerSimonian, R., & Freedman, D. S. (1992). Interpretation of linear regression models that include transformations or interaction terms. *Annals of Epidemiology, 2*(5):735-44.

Franzese, R., & Kam, C. (2009). *Modeling and interpreting interactive hypotheses in regression analysis*. Ann Arbor, MI: University of Michigan Press.

*Garcia, N. M., López, N., & Vélez, V. N. (2017). QuantCrit: rectifying quantitative methods through critical race theory. doi:10.1080/13613324.2017.1377675

*Gillborn, D., Warmington, P., & Demack, S. (2017). QuantCrit: Education, policy,'big data'and principles for a critical race theory of statistics. *Race Ethnicity and Education*, 1-22. doi: 10.1080/13613324.2017.1377417

Green, M. A., Evans, C. R., & Subramanian, S. V. (2017). Can intersectionality theory enrich population health research? *Social science & medicine, 178,* 214-216. doi:10.1016/j.socscimed.2017.02.029

Griffin, K. A., Bennett, J. C., & Harris, J. (2011). Analyzing gender differences in black faculty marginalization through a sequential mixed-methods design. *New Directions for Institutional Research, 2011*(151), 45-61. doi:10.1002/ir.398

Heckman, J. J., Urzua, S., & Vytlacil, E. (2006). Understanding Instrumental Variables in Models with Essential Heterogeneity. *Review of Economics and Statistics, 88,* 389-432.

Henseler, J., & Chin, W. W. (2010). A comparison of approaches for the analysis of interaction effects between latent variables using partial least squares path modeling. *Structural Equation Modeling, 17*(1), 82-109.

Hu, A., & Mustillo, S. A. (2016). Recent development of propensity score methods in observational studies: Multi-categorical treatment, causal mediation, and heterogeneity. *Current Sociology, 64*(1), 60-82. doi:10.1177/0011392115589599

Jann, B. (2013). *Predictive margins and marginal effects in Stata*. 11th German Stata Users Group Meeting, Potsdam (June 7, 2013). Retrieved from: https://www.stata.com/meeting/germany13/abstracts/materials/de13_jann.pdf

Kam, C. D., & Trussler, M. J. (2007). At the nexus of observational and experimental research: Theory, specification, and analysis of experiments with heterogeneous treatment effects. *Political Behavior*, 1-27. doi:10.1007/s11109-016-9379-z

Long, J. S., and J. Freese. 2006. *Regression models for categorical dependent variables using Stata, 2nd ed*. College Station, TX: Stata Press.

Manski, C. (2007). *Identification for prediction and decision*. Cambridge, MA: Harvard University Press.

Marshall, S. W. (2007). Power for tests of interaction: Effect of raising the Type I error rate. *Epidemiologic Perspectives & Innovations, 4*(1), 4.

*McCall, L. (2005). The complexity of intersectionality. *Signs: Journal of Women in Culture and Society, 30*(3), 1771-1800. Retrieved from http://www.gla.ac.uk/media/media_200317_en.pdf

Moffitt, R. (2008). Estimating marginal treatment effects in heterogeneous populations. *Annales d'Economie et de Statistique,* 239-261. doi:10.2307/27917247

Morabia, A. (2014). Interaction–Epidemiology's brinkmanship. *Epidemiologic Methods, 3*(1), 73-77. doi:10.1515/em-2014-0017

Morgan, S., & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research.* New York, NY: Cambridge University Press.

Museus, S. D., & Griffin, K. A. (2011). Mapping the margins in higher education: On the promise of intersectionality frameworks in research and discourse. *New Directions for Institutional Research, 2011*(151), 5-13. doi:10.1002/ir.395

Muthen, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology*, 267-316. doi:10.2307/271070

Overton, R. C. (2001). Moderated multiple regression for interactions involving categorical variables: A statistical control for heterogeneous variance across two groups. *Psychological methods, 6*(3), 218. doi:10.1037/1082-989X.6.3.218

Preacher, K. J., Curran, P. J., & Bauer, D. J. (2006). Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational and Behavioral Statistics, 31*(4), 437-448. doi:10.3102/10769986031004437

*Ragin, C. C., & Fiss, P. C. (2016). *Intersectional inequality: Race, class, test scores, and poverty.* Chicago, IL: Univesrity of Chicago Press.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Thousand Oaks, CA: Sage.

*Reid, R., Epstein, M. H., Pastor, D. A., & Ryser, G. R. (2000). Strengths-based assessment differences across students with LD and EBD. *Remedial and Special Education, 21*(6), 346-355. doi:10.1177/074193250002100604

Rhodes, W. (2010). Heterogeneous treatment effects: What does a regression estimate?. *Evaluation Review*, *34*(4), 334-361. doi:10.1177/0193841X10372890

*Schreiner, L. A., & Anderson, E. (2005). Strengths-based advising: A new lens for higher education. *NACADA Journal, 25*(2), 20-29. doi:10.12930/0271-9517-25.2.20

*Scott-Clayton, J., & Wen, Qiao. (2017). *Estimating returns to college attainment: Comparing survey and state administrative data based estimates.* CAPSEE working paper (January 2017). New York: Center for Analysis of Postsecondary Education and Employment. Retrieved from: https://capseecenter.org/estimating-returns-to-college-attainment/

Solon, G., Haider, S. J., & Wooldridge, J. M. (2015). What are we weighting for? *Journal of Human Resources, 50*(2), 301-316. doi:10.3386/w18859

Stata. (2017). Vce_options. (Online manual). Retrieved from: https://www.stata.com/manuals13/xtvce_options.pdf

*Thomas, S. L., & Heck, R. H. (2001). Analysis of large-scale secondary data in higher education research: Potential perils associated with complex sampling designs. *Research in Higher Education, 42*(5), 517-540. doi:10.1023/A:1011098109834

*Tourangeau, K., Nord, C., Le, T., Sorongon, A. G., Najarian, M., & Hausken, E. G. (2009). Early childhood longitudinal study, kindergarten class of 1998-99 (ECLS-K): Combined user's manual for the ECLS-K eighth-grade and K-8 full sample data files and electronic codebooks. Washington, DC: National Center for Education Statistics.

Turney, K. (2015). Beyond average effects: Incorporating heterogeneous treatment effects into family research. *Journal of Family Theory & Review*, *7*(4), 468-481. doi: 10.1111/jftr.12114

Vandenbrouke, J. P. (2013). The history of confounding. In A. Morabia (Ed.), *A history of epidemiologic methods and concepts* (pp. 313-326). Basel, Switzerland: Birkhäuser.

VanderWeele, T. J., & Knol, M. J. (2014a). Interactions and complexity: Goals and limitations. *Epidemiologic Methods, 3*(1), 79-81. doi:10.1515/em-2014-0016

VanderWeele, T. J., & Knol, M. J. (2014b). A tutorial on interaction. *Epidemiologic Methods, 3*(1), 33-72. doi:10.1515/em-2013-0005

Williams, R. (2012). Using the margins command to estimate and interpret adjusted predictions and marginal effects. *The Stata Journal, 12*(2), 308-331.

Winship, C., & Radbill, L. (1994). Sampling weights and regression analysis. *Sociological Methods & Research, 23*(2), 230-257. doi:10.1177/0049124194023002004

Wodtke, G. T., & Almirall, D. (2015). *Estimating heterogeneous causal effects with time-varying treatments and time-varying effect moderators: Structural nested mean models and regression-with-residuals.* Ann Arbor, MI: Michigan Population Studies Center.

Xie, Y. (2011). Causal inference and heterogeneity bias in social science. *Information Knowledge Systems Management*, *10*(1-4), 279-289. doi: 10.3233/IKS-2012-0197

Xie, Y. (2013). Population heterogeneity and causal inference. *Proceedings of the National Academy of Sciences*, *110*(16), 6262-6268. doi: 10.1073/pnas.1303102110

Xie, Y., Brand, J. E., & Jann, B. (2012). Estimating heterogeneous treatment effects with observational data. *Sociological Methodology, 42*(1), 314-347. doi:10.1177/0081175012452652